ECEN 250 Exam Review

On: May 4th, 10:30 AM (starts officially at 10:40 AM)

Requirements: Two cheat sheets (four sides, front and back), two pieces of scratch paper, a calculator, other stuff not allowed.

Medium: Canvas (online).

Contents: 80 questions in 75 minutes, some computations with a calculator, true/false & multiple choice, 35% is from before the midterm, 65% after the midterm, tested on speed.

Introduction:

- What is ML?
- What is learning? (types of learning)
- What we can learn from data

Data

- Types of data
- Structure of a dataframe
- Labeling
    - Supervised Learning needs labelling.
- Dealing with wrong, inconsistent, missing data.

Descriptive Statistics

- Population vs. Sample
- Measures of frequency
- Measures of central tendency
- Measures of dispersion
- Scaling
    - How to do various kinds of scaling, need this on your cheat sheet.

Clustering

- What is similarity?
    - Clustering is unsupervised, look for similarity amongst the data values.
- K-means clustering - basic understanding of phases and operations.
- What is K in K-means?
- Other clustering techniques.
    - Have a list of the other clustering techniques and characteristics associated with them.
- What tools do you have and what is the right tool for the job?

Probability

- What is a random variable, event (random variables to values), probability (tables)?

- We're dealing with discrete variables.
- What are distributions?
    - Joint, marginal, conditional
- Product rule & Bayes Rule
    - What is it and how to do it.
- Independence
    - How to show it.
- Expectation
    - What it is and how to calculate it.

Regression

- Linear - How do we fit data for linear regression?
    - Simple linear regression, first order polynomial (a line, $y = mx + b$).
- Error in fit
    - Re-read the error part of the regression lecture.
- Simple (single X), multiple (multiple Xs), and polynomial regression.
    - What's the difference between polynomial and linear regression?

Error, Underfit, Overfit

- Goodness of fit metrics.
- Bias vs. Variance.
    - What affects what in what direction?
    - Covered in the notes.
- Generalization
    - How well a model does on data it hasn't seen.
    - Very important than doing well off of data it's already seen.
- Regularization
- Covariance and Correlation of RVs.
    - Multivariate & bivariate statistics.
    - Know what the difference is and what we use.

Classification

- What's classification?
    - Predict to which class this observation will belong.
- K-Nearest Neighbors - General knowledge and when to apply.
- Accuracy, precision, recall, F1.
    - How do we derive these for binary classifiers in particular?

Inferencing

- How is Bayes' Rule used for Naive Bayes?
- Confusion matrix
    - What it is and how to use it.
- Know the general characteristics and parameters of a Naive Bayes Classifier.
    - Probabilistic model.

- What type of NB do we choose for different types of classification?
    - Different kinds for different kinds of models.
    - Characteristics of the Xs and their data types as well as the types of classes you expect to predict chooses your model.
    - Draw the matrix from the lecture notes.

Robust ML Methodology

- What is feature selection and engineering?
- Data Rules - Explicit for this class.
    - First Rule: Don't mess with your original data. Pull that data into a dataframe.
    - Second Rule: Don't test and train with the same data.
    - Third Rule: Don't validate with the same data you're training and/or testing.
        - Make decisions based on validation scores.
- Parameters vs. Hyperparameters
    - Parameters - Tune model.
    - Hyperparameters - Tune process.

Collinearity

- How do we determine collinearity and what do we do about it?
    - We determine by looking at a map of our correlations between random variables.
    - We want our independent variables (features) to be highly correlated with our output.
    - We don't want input features that're highly correlated.
    - Get rid of collinearities. Pick one or the other.
- If you see correlation greater than 0.8 or 0.9, that makes them linear and you need to know what to do about them (only for inputs).
    - We want inputs correlated with the output.

Parameters & Hyperparameters

- Parameters
    - Simple Regression: Slope & Intercept
    - Linear 2D classifier: Slope & Intercept
    - K-Means: Location of the Centroid
    - Artificial Neural Network: Weights
- Hyperparameters
    - K-Means: Number of Mean Points
    - K-NN: Number of Neighbors nearby that we vote on.
    - SVM (Support Vector Machines): Maximize the space between the red & green (margin).
    - Decision Tree: Tree Depth (how deep it goes).
    - Random Forest: Tree Depth & Number of Trees.
- What do parameters represent in Bayes' Classifier?
- Parameter estimation through counting.
- What to do for classes that haven't been seen?

    - Pretend that we've observed everything at least once.
    - Apply Laplace Smoothing.

Setting Probabilities

- Probability that a sign is a stop sign?
    - # of Observations of the Sign / Total # of Observations
    - COUNT OBSERVATIONS, NOT CLASSES

SVC

- Remember the Kernel Trick.
- What is a support vector?
- What's the difference between SV for classifier vs. regressor?
    - Difference between definitions.
    - Margin vs. E2.

Decision Trees

- Learning Types
- Phenomenon vs. True Function vs. Approximate Function
- How decision splits are made.
    - What metrics do we use to determine these splits?
    - Feature importance?
- General knowledge on when to use DTs.

Random Forest

- What is overfitting?
    - How do we detect it?
    - How do we minimize it?
- What's an ensemble?
- Weak learner
    - Relatively inaccurate model.
- Strong learner
    - Ensemble of weak learners.
- Boosting, bagging, pasting.

Choosing the Best Classifier Model

- Balances vs. Imbalanced?

MIDTERM

Neural Networks

- What are the parts in an artificial neuron?
    - Common activation functions.
- What are the parameters in an artificial neural network?
- How do we size them?

    - Inputs & Outputs
- Hidden Layers
    - Number of hidden layers to create a deep neural network/

CNNs

- Filters.
- Look for characteristics in our spacial features.
- Convolutions Layers and Pooling Layers.

CNNs for Images

- What is object classification?
- What is object detection, what two functions does it perform, and what methods are used?
- What is segmentation and what methods are used to accomplish this?
- What is transfer learning?

RNNs

- Added feedback into neurons.
    - Inputs into the summing network.
    - We have memory.
- Change the way that we train them and set the wait.
- Encoder & Decoder
- LSTM - Long & Short Term Memory.

Signals

- Sequences, series, signal, and time series.

Transformers

- Add in extra complexity and functions into our model.
    - You should know what these are.

Limits

- Correlation Characteristics
- Causality
    - Add some extra tests for us to see if A caused B.
- Data issues limiting ML.
- Human failures limiting ML.